Visual Analytics for Biomedical Cluster Subdivision: A Design Study with Psychiatrists

Jihye Lee, Hyoji Ha, Hyunwoo Han, Sungyun Bae
Lifemedia Interdisciplinary Program
Ajou University
South Korea
alice0428, hjha0508, ainatsumi, loah @ajou.ac.kr

Sangjoon Son, Changhyung Hong
Department of Psychiatry,
School of Medicine, Ajou University
South Korea
sjsonpsy, antiaging@ajou.ac.kr

Hyunjung Shin
Department of Industrial Engineering
Ajou University
South Korea
shin@ajou.ac.kr

Kyungwon Lee
Department of Digital Media
Ajou University
South Korea
kwlee@ajou.ac.kr

ABSTRACT

In the last few years, Electronic Health Records (EHRs) have collected large-sized medical data. While EHR allows doctors to approach medication data easily, they suffer difficulties to analyze multi-dimensional medical data. We present multi-dimensional visual analytics tools to support analyzing multi-dimensional dataset by the combination of 3D RadVis and parallel coordinate. Also, we propose user-driven research design process to prospect for visualization development. This study is now in progress to interview domain experts to analyze the usability of the tools.

CCS CONCEPTS

• **Human-centered computing** → **Visualization**; *Visualization application domains*; Visual analytics • **Human computer interaction (HCI)** → *HCI design and evaluation methods;* User studies

KEYWORDS

Visual Analytics, User study, Information visualization, Health Informatics.

CGI '17, June 27-30, 2017, Yokohama, Japan

ACM ISBN 978-1-4503-5228-4/17/06...$15.00
http://dx.doi.org/10.1145/3095140.3095157

1 INTRODUCTION

By adopting a digital medical record management architecture, the EHR (Electronic Health Record) system, doctors realize that analyzing EHR data is crucial to conduct reliable medical research. Medical experts start to prospect disease patterns for efficient medication by collaborating with data scientists. For example, LifeLines presented a case study to show the medication patterns of each patient [15, 19]. Wang et al. presented chart visualization to investigate chronic disease progression [19]. TimeSpan described hospital timeline data to explore stroke treatment process. This approach helped medical expert to understand stroke treatment process more in detail [12]. Also, clustering analysis are used in psychiatry studies. It has been adopted for a long time to distinguish disease group. For instance, psychologists used Ward's method, a kind of cluster analysis, to understand why young women suffer from eating disorders [5]. In recent studies, Petrovic et al. [14] suggest the dementia patient subdivision by Spearman's correlation based on Neuropsychiatric inventory. In this paper, we suggest a design study that engaged with psychiatrists. We collaborated with two psychiatrists as participants and co-authors of this paper to help the process. In addition, this study mainly refers the design framework suggested by Sedlmair et al. [16] and Sohaib Ghani et al. [7] Based on the literature, we considered the demands before developing visual analytics tool. For this, we followed design process that included (1) understanding the demands from domain experts when they analyze multi-dimensional healthcare data (2) defining the problems based on the relevant works. (3) designing visualization, (4) visualization implementation and (5) qualitative evaluation with domain experts. Due to the lack of multi-dimensional data analysis experience of domain experts, we suggested the patient subdivision case study to clarify the function of visual analytics tool. Our contributions in this study is following: (1) subdividing

Alzheimer's disease patient group (2) proposing a visual analytics tool to support decision making from domain experts. (3) proposing a methodology to prevent node overlapping. Finally, we hope to clarify that this study focused on problem-driven and user-driven. Therefore, technical novelty is not the goal until the complete this project. This study is now in progress on interviewing domain experts and partial research process is written in the paper.

2 RELATED WORKS

2.1 Cluster Analysis

Cluster analysis uses various academic disciplines: psychiatry, psychology and social science etc. This methodology aims to divide data into diverse groups that have different features. In the visualization field, Ankerst et al. [1] developed a visualization method to show multi-dimensional data after hierarchical clustering. While if a user investigates the characteristics of the multi-dimensional data, it is hard to draw parallel lines. Beham et al. f[2] suggests a methodology that combines a radial tree map and Parallel Coordinates. It shows a radial tree map distribution, through controlling all the axes of Parallel Coordinates.

2.2 Parallel Coordinate

Parallel Coordinate Plots (PCPs) [11] are a popular way to visualize the distributions and ranges of multi-dimensional data. The parallel axes mean the vertical bars represent variables, while graph lines draw the pattern of each objective. However, it has limitations, in that it is not able to present the Pareto front shape. Despite this weakness, PCP is used generally to visualize multi-dimensional objects. For example, Fua et al. [6] shows parallel visualization based on hierarchical clustering. This represents information clusters of cluster data. In another case, Zhou et al. [20] proposes visual clustering methods that apply curved line and visual bundle technology in Parallel Coordinates. These methods help to observe Parallel Coordinate lines in more detail and from various perspectives.

2.3 RadVis

Radial coordinate visualization (RadVis) [9] uses a nonlinear system to locate the nodes from n-dimensional points to a 2-dimensional map. Enrico Bertini [3] suggests the methodology that combines RadVis and Parallel Coordinates. In this case, each cluster has a specific color. Also, they developed Parallel Coordinates to select the specific area to choose the node in RadVis. We were inspired by the framework of a visualization system in the research of Ibrahim et al. [10], who suggests 3D RadVis using Pareto front methods to make a spectrum of node distributions. In this study, we suggest a multi-dimensional data visualization tool based on the needs of medical experts. For this reason, we combined 3D RadVis and Parallel Coordinates to visualize the Electronic Healthcare Records (EHRs) of dementia patient cohorts.

3 RESEARCH OVERVIEW

The goal of this study is subdividing dementia patient groups to support psychiatrists' data analysis. They hope to characterize dementia patient group which has the diverse spectrum. In order to figure out the possibilities of the patient subdivision, visualization researchers planned to develop visual analytics tool. We decided to adopt a user-driven design process with the following stages: (1) understanding the demands from domain experts (2) defining the problem (3) designing visualization (4) visualization implementation (5) second qualitative evaluation with doctor and clinical psychologist. We add the dimension which represents variable range and score, and develop 3D visualization. This approach contributes to solve the node duplication in 2D RadVis.

3.1 Casting

We met two psychiatrists to understand the demands about dementia patient clustering. First, they hope to clarify how to subdivide dementia patient group based on statistical approach. Dementia is the cognitive disorder which is ambiguous to diagnose. Despite its elusiveness, to cure dementia, doctors should target the use of drugs to treat mental disorder. Second, psychiatrists hope to compare the scores of divided dementia clusters. For example, domain experts want to compare high-risk and low-risk Alzheimer's disease group. In conclusion, through the interview with the domain experts, they hope to divide patient groups separately, to define and analyze dementia in more detail. As the result of the discussion, we collected the needs from domain experts that subdivide the patient groups separately.

3.2 Discover

In this stage, we tried to define the needs from domain experts, and reviewed existing works of visualization field and medical science. After reviewing backgrounds, we set the guidelines mentioned in the related work [5, 18]. This literature proposes several representatives for clustering analysis. Based on our reviews, we developed the design guidelines below, the list of notions to specify clusters. We follow the measurement guidelines of cluster analysis while developing the visual analytics tool.

(1) Representativeness of cluster: How well do specific nodes represent clusters?

(2) Efficiently finding nearest neighbors: What does the similarity between distances mean? What does the nodes in one cluster maintaining a constant distance mean?

(3) Segmentation and partitioning: How can we know that a cluster is well segmented?

Performing cluster analysis on visual analytics tool is the most important demand to psychiatrists. Therefore, we considered the approach to subdivide the patient group by cognitive assessment score in the Clinical Research Center for Dementia of South Korea (CREDOS) [4].

3.2.1 Dataset We used the dementia examination cohort data named CREDOS [4], which has 21,094 electronic health records. It includes variables measuring cognitive functions to diagnose patients. Table 1. below explains CREDOS. This cohort was collected by 37 hospitals in South Korea from 2005 to 2013. After refining the data by ruling out the cases with missing factors, we selected 2,219 records via stratified sampling.

Table 1. Data components included in CREDOS.

Variables	Explanation
Patient information	Cohort ID, Personal information (gender, age, educational background), Physical examination
Caregiver information	Caregiver's information (gender, age, educational background, relationship between patient and caregiver)
Cognitive assessments	Caregiver-Administered Neuropsychiatric Inventory (CGA-NPI), Seoul-Instrumental Activities of Daily Living (S-IADL), Diagnosed disease (SMI, MCI, VCI, SVD, AD)

CREDOS includes five dementia stages, from Subjective Memory Impairment (SMI) to Alzheimer's Disease (AD). It is divided according to the chronicity. First, SMI is the weakest stage of dementia. Second, Mild Cognitive Impairment (MCI) is the decisive point to move to the chronic stage. Third, Vascular Cognitive Impairment (VCI) and Subcortical Vascular Dementia (SVD) include the patients who have both stroke disease and cognitive disorder. The final stage, Alzheimer's Disease (AD), is famous as the stereotype of dementia. At this time, we focus on subdividing the Alzheimer's disease patient group which is most widely distributed node group.

3.3 DESIGN

3.3.1 First visualization design At the First time, we designed a 2D node-link diagram visualization to explore the relationships between each health record. Nevertheless, when we visualized EHR data with the node-link diagram, it had limitations in presenting EHR data. Due to the dimensional limitation of 2D node-link diagram, it is not proper model to visualize multi-dimensional data. Fig. 1 shows node distribution of 2D node-link diagram and 3D RadVis. For example, green colored group (SMI and MCI groups) in the 2D node-link diagram appears as the same cluster. However, in 3D visualization, subdivision result shows the distinction of node distribution. Though SMI and MCI have similar symptoms in clinical field, they have diagnosed different name by MRI (Magnetic Resonance Imaging) result. Nevertheless, the subdivision result in 3D RadVis clearly shows node distribution difference. When we showed the first prototype to the psychiatrists, as a co-author, they consented the good point of data visualization. however, domain expert asked a question how to subdivide each patient group in the tool. 2D node-link diagram has the problem that expressed same node position if the nodes have the same total score in the medical examination. Sometimes dementia patients can have the same score though they are differently diagnosed. Therefore, we should solve the problem when each medical record, the node, is duplicated in the visualization. This point made us decide to change visualization format.

3.3.3 Second visualization design After observing the difference between 2D and 3D visualization, we considered to add a dimension to the visual analytics tool. We inspired visualization system in the research of Ibrahim et al. [10], who suggests 3D RadVis using Pareto front methods to make a spectrum of node distributions. In this case, three-dimension to understand node distribution in detail using height [17]. Moreover, this approach supports to prevent node overlapping in the visualization. Even if they are assigned the same space in 3D RadVis, sometimes each node scores can be different. To compensate for this defect in 3D RadVis, we adopt dimensional anchor that makes node position based on average score. The dimensional anchor is located the vertex of the intersection point from the bottom side to a vertical line. This made the node position to the intermediate point by pulling a node in the highest or lowest point of each variable. It prevents node duplication in 3D RadVis and collocates the node position properly. [8]

3.4 IMPLEMENTATION

After visualization design process, we developed visual analytics tool combined 3D RadVis and Parallel Coordinates. This visual analytics tool can interact between 3D RadVis and Parallel Coordinates. This tool supports repetitive clustering even though user already split the clusters. If they hope to subdivide cluster more, they can choice the methodology between k-means and forgy algorithm. In addition, the user can rotate the 3D RadVis cylinder to observe the data distribution.

This tool has several interactions. First, when the user selects a node in 3D RadVis, the graph lines of each node will appear in Parallel Coordinates. Also, we apply multi-filtering to adjust the axes in Parallel Coordinates. This helps to make multi-range selections between variables. If the user hopes to select duplicate score ranges in the same axis, they can drag the axis many times. This interaction was suggested by psychiatrists to compare the score ranges for each variable. Also, users can export the selected cluster data as CSV dataset from the tool. It helps to analyze data with another data analysis program. This function was also suggested by our two domain experts, and they believe that this will be useful to target each disease patient group, so that doctors can develop drugs that are suitable for specific patient groups.

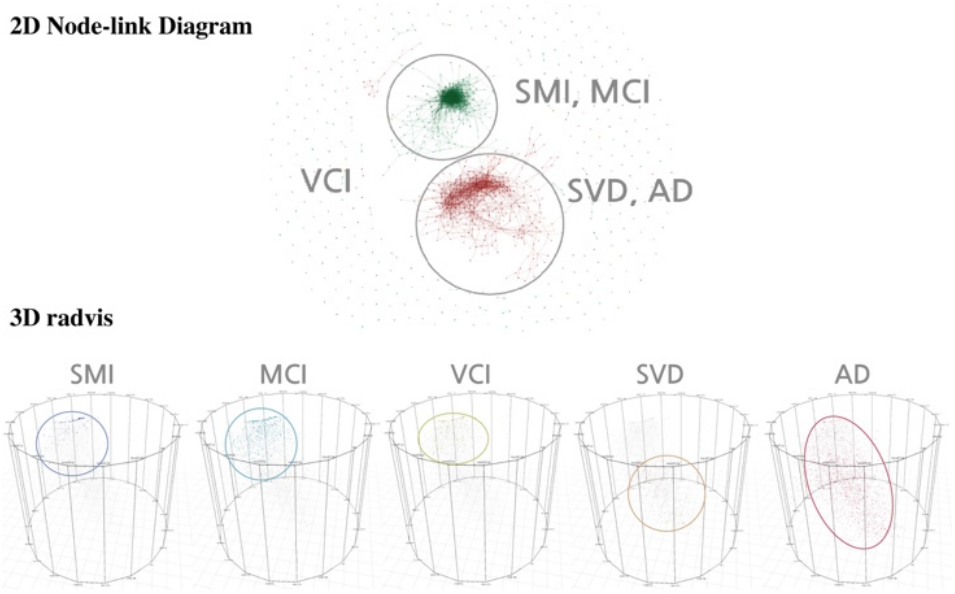

Figure 1: A Visualization analysis system to subdivide dementia diagnosis stages: 3D RadVis and Parallel Coordinates presented in the green box on the left suggest variables in physical examinations on individual patients while the information in the red box on the right was based on variables in cognitive assessments.

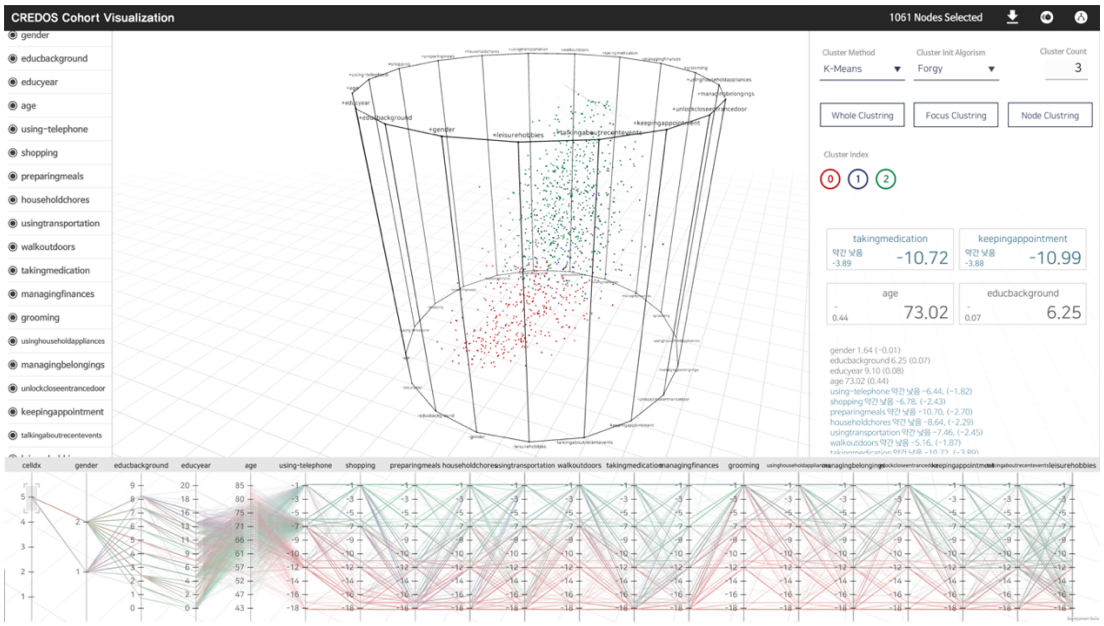

Figure 2: Visualization combined 3D RadVis and Parallel Coordinates: This visual analytics tool supports multi-dimensional data analysis. The user can control the RadVis cylinder and parallel graph. Also, they can divide and explore the data various ways to use additional functions in visual analytics tool.

Fig. 2 shows the entire interface of visual analytics tool. The left side of the tool is variables list in raw data. Users can choice the variables they hope to explore. The right side of Fig. 2 shows the information to figure out the characteristic of each cluster.

3.5 CASE STUDY

We conduct a case study based on questionnaires of the Seoul-Instrumental Daily Living (S-IADL) test. This test examines daily behaviors such as shopping, taking medication, and using transportation. By suggestion from two psychiatrists, we subdivided dementia patient group with 15 variables of S-IADL test. Domain experts are interested in detecting dementia patients based on daily behavior. As a result, we can subdivide the clusters between Alzheimer's disease patient groups. Bottom of Fig. 1 shows the separated dementia group distributions in 3D visualization. Especially, AD group shows the separation from low-risk factor to high-risk factor group. Fig 3 shows the variables subdividing AD group. The high-risk factors of Alzheimer's disease include the variables such as shopping, using transportation, walking outdoors, and unlocking doors. In particular, the low-risk factor cluster has the variables including taking medication, managing finances, keeping appointments, and talking about recent events. In all groups, Alzheimer's disease patients received low scores in examining personal relationships. We discussed this case study result with the psychiatrists. They agreed with the result that human relationship factors affect disease progression. Also, we understand Alzheimer's disease can progress separately. Consequently, we realized that daily behaviors influence the progression of dementia. We could partially verify design guidelines mentioned discover stage by result of this case study; (1) Representativeness of cluster. It can observe from the score difference between each group. They have 5 to 7 score difference between high-risk and low-risk group. Also, the lowest score differences are shown from 3 to 5 between the middle-risk and low-risk groups.

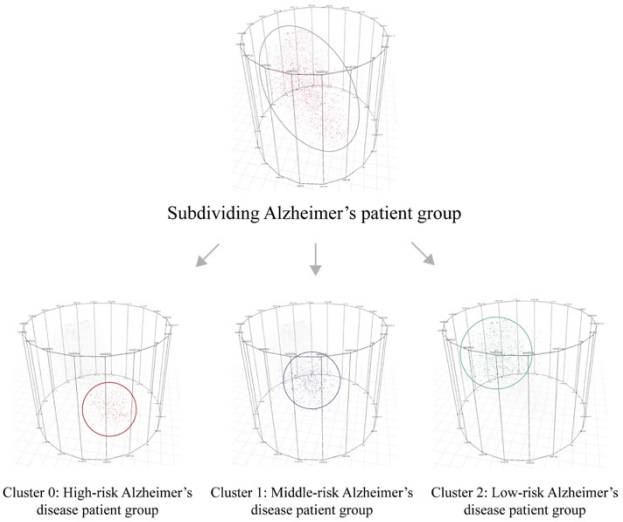

Figure 3: Cluster subdivision result of Alzheimer's disease group. It divided 3 segmentations by behavior difference based on S-IADL test result.

4 CONCLUSION

This study aims to suggest a visual analytics tool to support cluster subdivision for domain experts. Based on the relevant works, 3D RadVis, Parallel Coordinates, and cluster analysis, we developed a visualization tool that combined 3D RadVis and Parallel Coordinates. By supplement the dimension to visualize multi-dimensional data, it aids to subdivide dementia patient clusters. We conducted a case study based on a daily living questionnaire. This study contributes to doctors understanding biomedical data through a visual approach. Also, we challenged multi-dimensional EHR data. We expect that doctors will use this tool to subdivide dementia patient clusters. Though we could not interview more experts (e.g. clinical psychologists, neurologists), this research is now in progress of interviewing more to understand the usability of the tool. Also, we are now in progress to verify all design guidelines in discover stage.

ACKNOWLEDGEMENT

This project was supported by a National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) (No.2015R1 A5A7037630), the Korea Healthcare Technology Research and Development Project, Ministry of Health & Welfare, Republic of Korea (HI10C2020), and under 2016 BK21 Program by Ajou University. Also, we thank to Seongkyeong Lee and Joonho Yeo for constructive feedback.

REFERENCES

[1] M. Ankerst, S. Berchtold, & D. A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In proceedings of Information Visualization, (VIS 1998), North Carolina, United States, 52-60, DOI: 10.1109/INFVIS.1998.729559

[2] M. Beham, W. Herzner, M. E. Gröller, & J Kehrer, Cupid: cluster-based exploration of geometry generators with parallel coordinates and radial trees. IEEE transactions on visualization and computer graphics, 20,12(2014), 1693-1702, DOI: 10.1109/TVCG.2014.2346626

[3] E. Bertini, L. Dell'Aquila, G. Santucci. 2005. Springview: Cooperation of radviz and parallel coordinates for view optimization and clutter reduction. In Proceeding of the Coordinated and Multiple Views in Exploratory Visualization (CMV 2005). IEEE, 22-29. DOI: 10.1109/CMV.2005.17

[4] S,H, Choi, J. Lee, S.J. Kim, J.Y. Choi, J.W. Kwon, B.N. Yoon, Y.S. Yang, S.Y Kim, J.H Jeong. 2014. Driving in Patients with Dementia: A CREDOS (Clinical Research Center for Dementia of South Korea) Study. *Dementia and Neurocognitive Disorders*, 13, 4 (2014), 83-88.

[5] B. Everi , S. Landau, S. Leese, D. Stahl, 2011, Cluster Analysis, 5th Edition, 71-110. WILEY SERIES IN PROBABILITY AND STATISTICS, New Jersey, 1.5.3. Psychiatry.

[6] Y. H. Fua, M. O. Ward, & E. A. Rundensteiner, Hierarchical parallel coordinates for exploration of large datasets. In Proceedings of the conference on Visualization'99 (VIS 1999), California, United States, 43-50. DOI: 10.1109/VISUAL.1999.809866

[7] S Ghani, B. C. Kwon, S. Lee, J. S. Yi, & N. Elmqvist, Visual analytics for multimodal social network analysis: A design study with social scientists. IEEE Transactions on Visualization and Computer Graphics, 19,12, (2013), 2032-2041, DOI: 10.1109/TVCG.2013.223

[8] H. Ha, H. Han, S. Bae, J. Lee, S. Son, C. Hong, H. Shin, K. Lee, A Study on Visualization Methods of Semantic Clustering for Multidimensional data, Communications of the Korean Institute of Information Scientists and Engineers, 11,34, (2016),51-61.

[9] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, & E. Stanley, DNA visual and analytic data mining. In Proceedings of Visualization'97(VIS 1997), Arizona, United States, 437-441. DOI: 10.1109/VISUAL.1997.663916

[10] A. Ibrahim, S. Rahnamayan, M. V. Martin, & Deb, K. Deb, 3D-RadVis: Visualization of Pareto front in many-objective optimization. In proceding of *Evolutionary Computation (CEC 2016), IEEE, Vancouver, Canada*, 736-745. DOI: 10.1109/CEC.2016.7743865

[11] A. Inselberg. 1985. The plane with parallel coordinates. *The visual computer*, 1,2 (1985), 69-91.

[12] A. Lhuillier, C. Hurter, C. Jouffrais, E. Barbeau, HL Amieva. 2015. Visual analytics for the interpretation of fluency tests during Alzheimer evaluation. In *Proceedings of the 2015 Workshop on Visual Analytics in Healthcare* (VAHC '15). ACM, New York, NY, USA, Article 3, 8 pages. DOI: https://doi.org/10.1145/2836034.2836037.

[13] M.H. Loorak, C. Perin, N. Kamal, M. Hill, S. Carpendale. 2016. TimeSpan: Using visualization to explore temporal multi-dimensional data of stroke patients. *IEEE transactions on visualization and computer graphics*, 22,1(2016), 409-418. DOI: 10.1109/TVCG.2015.2467325

[14] M. Petrovic, C. Hurt, D. Collins, A. Burns, V. Camus, R. Liperoti, A. Marriott, F. Nobili, P. Robert, M. Tsolaki, B. Vellas, F. Verhey, E.J. Byrne, Clustering of behavioural and psychological symptoms in dementia (BPSD): a European Alzheimer's disease consortium (EADC) study. *Acta Clinica Belgica*, *62*, 6 (2008), 426-432. DOI: http://dx.doi.org/10.1179/acb.2007.062

[15] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, & B. Shneiderman, (1998). LifeLines: using visualization to enhance navigation and analysis of patient records. In Proceedings of the AMIA Symposium (AMIA 1998). American Medical Informatics Association. Florida, United States, 76

[16] M. Sedlmair., M. Meyer, & T. Munzner, Design study methodology: Reflections from the trenches and the stacks. IEEE Transactions on Visualization and Computer Graphics, 18,12(2012), 2431-2440. DOI: 10.1109/TVCG.2012.213

[17] JH Sung, DY Lee, HK Kim, Difference of GUI Efficiency based on 3D and 2D Graphic -Imaginary 3D IPTV Interface Development Using Virtual Reality Theory-, Journal of Content, 7,7(2007), 87-95.

[18] P.N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining 1st Edition, 487-568, Addison-Wesley Longman Publishing Co., Inc. Boston, Chapter 8 Cluster Analysis: Basic Concepts and Algorithm

[19] C.F. Wang, J. Li, K.L. Ma, C.W. Huang, Y.C. Li. 2014. A visual analysis approach to cohort study of electronic patient records. In *Proceeding of the Bioinformatics and Biomedicine* (BIBM 2014). IEEE, Belfast, UK, 521-528, DOI: 10.1109/BIBM.2014.6999214

[20] H. Zhou, X. Yuan, H. Qu, W. Cui, B. Chen. 2008. Visual clustering in parallel coordinates. Computer Graphics Forum, 27, 3 (2008), 1047-1054. DOI: 10.1111/j.1467-8659.2008.01241.